Ejercicio 1

a) Implementado en notebook. Usa método .fit de clase SGD.

b) Hay 9 parámetros desconocidos

- $w_{1,1}^{(1)}$
- $w_{1,2}^{(1)}$
- $b_1^{(1)}$

} neurona 1

- $w_{2,1}^{(1)}$
- $w_{2,2}^{(1)}$
- $b_2^{(1)}$

} neurona 2

} capa 1

- $w_{1,1}^{(2)}$
- $w_{1,2}^{(2)}$
- $b_1^{(2)}$

} neurona 1

} capa 2

c) Hiperparametros

- Número de epochs (probamos por ejemplo 1, 10, 100, 1000)
- Learning rate [lr] (podemos probar por ejemplo 0.1, 0.01, 0.001)
  - Al variar el learning rate variamos la velocidad de convergencia a nuestro mínimo.
    Si el lr es pequeño necesito más epochs dado que el cambio que produce el lr en nuestro pesos es menor. Si es grande la convergencia es más rápida (menos epochs). El problema Si es muy grande es que puede converger a una solución subóptima. y si es muy pequeño puede hacer que en un punto no podamos mejorar más en el proceso de llegar a un mínimo (nos "atascamos").
- La arquitectura de nuestra red también son hiperparámetros (cantidad de layers y de neurons por layer). En nuestro caso ya están prefijados y no los podemos cambiar.

d) implementado en metodo predict.

Derivadas calculadas

capa 2

$$\frac{\partial L}{\partial w_{i,1}^{(2)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{dw_{i,1}^{(2)}} = -2(y_i - \hat{y}_i) \cdot a_{i,1}^{(1)}$$

$$\frac{\partial L}{\partial w_{i,2}^{(2)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{dw_{i,2}^{(2)}} = -2(y_i - \hat{y}_i) \cdot a_{i,2}^{(1)}$$

$$\frac{\partial L}{\partial b_1^{(2)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{db_1^{(2)}} = \underbrace{-2(y_i - \hat{y}_i)}_{err_i} = -2\, err_i \qquad \sigma(x) = \frac{1}{1+e^{-x}}$$

capa 1

$$\frac{\partial L}{\partial w_{i,1}^{(1)}} = \frac{dL}{d\hat{y}} \cdot \frac{d\hat{y}}{da_{i,1}^{(1)}} \cdot \frac{\partial a_{i,1}^{(1)}}{\partial z_{i,1}^{(1)}} \cdot \frac{\partial z_{i,1}^{(1)}}{\partial w_{1,1}^{(1)}} = -2 \cdot err_i \cdot w_{1,1}^{(2)} \cdot \sigma(z_{i,1}^{(1)}) \cdot (1 - \sigma(z_{i,1}^{(1)})) \cdot x_{i,1}$$

$$\frac{\partial L}{\partial w_{i,2}^{(1)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{da_{i,1}^{(1)}} \cdot \frac{da_{i,1}^{(1)}}{\partial z_{i,1}^{(1)}} \cdot \frac{\partial z_{i,1}^{(1)}}{\partial w_{1,2}^{(1)}} = -2 \cdot err_i \cdot w_{1,1}^{(2)} \cdot \sigma(z_{i,1}^{(1)}) \cdot (1 - \sigma(z_{i,1}^{(1)})) \cdot x_{i,2}$$

$$\frac{\partial L}{\partial b_1^{(1)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{da_{i,1}^{(1)}} \cdot \frac{da_{i,1}^{(1)}}{\partial z_{i,1}^{(1)}} \cdot \frac{\partial z_{i,1}^{(1)}}{\partial b_1^{(1)}} = -2 err_i \cdot w_{1,1}^{(2)} \cdot \sigma(z_{i,1}^{(1)}) \cdot (1 - \sigma(z_{i,1}^{(1)})) \cdot 1$$

$$\frac{\partial L}{\partial w_{2,1}^{(1)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{da_{i,2}^{(1)}} \cdot \frac{da_{i,2}^{(1)}}{\partial z_{i,2}^{(1)}} \cdot \frac{\partial (z_{i,2}^{(1)})}{\partial w_{2,1}^{(1)}} = -2 err_i \cdot w_{1,2}^{(2)} \cdot \sigma(z_{i,2}^{(1)}) \cdot (1 - \sigma(z_{i,2}^{(1)})) \cdot x_{i,1}$$

$$\frac{\partial L}{\partial w_{2,2}^{(1)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{da_{i,2}^{(1)}} \cdot \frac{da_{i,2}^{(1)}}{\partial z_{i,2}^{(1)}} \cdot \frac{\partial z_{i,2}^{(1)}}{\partial w_{2,2}^{(1)}} = -2 err_i \cdot w_{1,2}^{(2)} \cdot \sigma(z_{i,2}^{(1)}) \cdot (1 - \sigma(z_{i,2}^{(1)})) \cdot x_{i,2}$$

$$\frac{\partial L}{\partial b_2^{(1)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{d\hat{y}}{da_{i,2}^{(1)}} \cdot \frac{da_{i,2}^{(1)}}{\partial z_{i,2}^{(1)}} \cdot \frac{\partial z_{i,2}^{(1)}}{\partial b_2^{(1)}} = -2 err_i \cdot w_{1,2}^{(2)} \cdot \sigma(z_{i,2}^{(1)}) \cdot (1 - \sigma(z_{i,2}^{(1)})) \cdot 1$$